JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-10, Oct 2019 E-ISSN: 2347-2693

Dengue Prediction Using Tweets in India

Sarita Kumari[1*], K. Jeberson[2], W. Jeberson[3]

[1, 2, 3]Dept of Computer Science and Information Technology, Vaugh Institute of Agricultural Engineering and Technology,
Sam Higginbottom University of Agriculture, Technology & Sciences, Allahabad, Uttar Pradesh, India

**Corresponding Author: ksarita232@gmail.com, Tel.: +919264989825*

DOI: https://doi.org/10.26438/ijcse/v7i10.5763 | Available online at: www.ijcseonline.org

Accepted: 12/Oct/2019, Published: 31/Oct/2019

Abstract— In India, people have started using twitter and nowadays, its craze has overshadowed the users all day. In India, a Twitter user across India was predicted to be more than 34 million in 2019. Twitter data is a very huge amount of data that can be used for the prediction of various diseases. Tweets are strongly related to Dengue cases. Dengue is a viral-borne disease that is also one of the widespread waterborne diseases. Nowadays people are trying a lot to avoid being a victim of dengue. But this communicable disease has highly increased alongside the urbanization rate in the tropical rain forest region. In this research paper, we focused on the retrieval of tweets using a hashtag keyword using a free analytic tool Vicinitas. We collected a set of 102 tweets to train a classifier to identify dengue, record and predict the emergence and transmission of dengue in a population. WEKA is a collection-set for machine learning and it is free open-source software. In this research, we used the dengue datasets with a total of one hundred two instances of dengue and two attributes i.e., text and class to determine accuracy using the various classifying algorithms. For the best outcome, we used seven classification techniques for accuracy. The main methodology and the techniques we used for predicting the dengue are J48, Naïve Bayesian, SMO, and Random tree, ZeroR, Random Forest and REP Tree. We after evaluating various attributes of the result finally concluded that Bayes obtained the highest accuracy rate.

***Keywords*—**Dengue, Weka, and Classification.

I. INTRODUCTION

Dengue is an arboviral contamination caused by mosquitoes. It is primarily a mosquito-borne infection that causes severe flu and dengue risk. The female mosquitoes live near humans in subtropical and tropical regions. In our country, India, confirmed Dengue cases are mainly reported in the urban areas. India is endemically to dengue. Dengue is spread widely on an average in the month of April to September due to heavy rainfall i.e., transmits during the months of April to November from north to the south [1]. After feeding on an infected person's blood containing the dengue virus, the female Aedes aegypti becomes the dengue vector. The syndrome of dengue is low blood pressure, headache, muscular joint pain, metallic taste and rashes [2]. Dengue is a cyclical type of an infected body, present inside the skin. There is no antibiotic available for this treatment. In the cycle mechanism by which dengue is transmitted, it is spread through the saliva of Aedes mosquito.

We took data from social media networks like Twitter for our study in avoiding big diseases like dengue. Today's young generation highly prefer a socialized platform to share their thoughts and feelings about various topics, including healthcare-related topics that are widely spread in our society. Twitter is a real-time social networking service that has gained large popularity. It is an online broadcasting channel that continues as a distinct form of blogging. It is also called a microblogging site, on which users update the post and collaborate with information, a directive letter known as tweets. The tweets were finite to 140 words; however on 7 November 2017 the limit was increased to 280 for all languages. More than 300 million operating users around the world posts around 600 million tweets daily, including re-tweets. There are about 34 million engaged users in India, which means that nowadays it is becoming very popular in our country too. This is helpful in detecting any form of the disease, as people are taking an interest in updating their daily life by sharing their activities as thoughts, pictures, videos, and gifs, etc [3].

We collected datasets from a twitter tool called Vicinitas on dengue. Dataset for dengue gives information about the patient suffering from dengue disease. We extracted tweets that had either of the following keyword 'Dengue' and hashtag '#Dengue' from a period of January 2018 to January 2019. Tweets can be thought of as a mixture of structured data and unstructured data. The collected Excel file contains data such as the user's name, followers, language, time of the day, the time zone, the geographical location, etc. Given such a structure of data, it becomes necessary to devise an effective and real-time solution to extract keywords from the tweets in such a way that the meaning of tweets is not compromised. Once this careful extraction is successfully performed, we can then use the tweets for analysis. We collected more than 2000 tweets from India but we selected 102 tweets to classify the dengue fever. The main purpose of this study is:

1. To classify the useful Data for classification of dengue disease.
2. To compared text mining or data mining algorithms on collective datasets.
3. To identify the best techniques using algorithms for predicting the dengue dataset.

Text Mining acts as the process of exploring a large amount of text data. Also, it is used to analyze a vast number of unstructured datasets. It is aided by software that can identify keywords and attributes in the data. It is the alias to Text Analytics. Text Mining is similar to data mining were data mining problems are solved by Weka software. Weka is best for solving problems in the bioinformatics research area. Weka developed by the Waikato University of Hamilton, New Zealand was implemented in 1997 is freely available on the web. It is a java-based language that is portable in every developed system. Weka has five interfaces to classify and predict the accuracy of the dengue disease. It has several types of processing tools and different types of algorithms are used to produce an accurate result, many nominal clustering structures, and associative rules. From the dengue dataset, weka algorithms are applied for generating accurate results by extract meaningful data. The objective of the paper is disease prognosis using a data mining tool and then determines which algorithm is best for dengue analysis. All the algorithms are compared with each other to give the best accuracy result. The following techniques were implemented in the stored dataset for determining the accuracy, classification, and comparison of results.

Several Data Mining Techniques are used for the prediction of Dengue Fever. The technique is a systematic approach to build a model from the input dataset. Classification approaches are much suited for the analysis of data address with binary classes of nominal categories. The techniques include algorithms such as Naïve Bayes, REP tree, SMO, Zero R, Random Forest, J48, Random Tree, etc. These techniques employ an algorithm to recognize an idea that best matches the link betwixt the attribute set and the class label of the dataset. These algorithms have good capability i.e., that accurately predicts the class of previously anonymous records. The paper is systematized as follows, Section I: Introduction contains a full description of this research that how we are taking dengue as a serious disease and using tweets for faster predictions. Section II Related Work: This section shows the main related work that is the classification of the dengue dataset by using weka and their result suite to predict dengue. Section III Methodology: shows how we used the weka tool on the dengue dataset after training it, Section IV Result and Discussion: how we found out the best result and discussion about it. Section V Conclusion: finally we conclude the best outcomes of this research paper and how we proceed with this research in a better way in the future.

II. LITERATURE REVIEW

Wajeeha Farooqi, Sadaf Ali, used the datasets from different hospitals. They applied data mining techniques to classify the Dengue Haemorrhagic fever (DHF) along with Dengue fever. The researcher used the Classification Algorithms which calculated their performance based on the accuracy, sensitivity, specificity, precision, and false-negative rate. They used techniques like Naïve Bayes (NBC), Support Vector Machine, K-nearest neighbor (KNN), Decision Tree (DT), and Multilayer Preception. Among all of them, the multilayered perception and decision tree classifiers have the best ability to correctly identify the DHF and it also has the lowest false-negative rate from the dataset. Next, the Naïve Bayes identified the highest acuteness and the best 14 attributes [2].

N. Saravanan, Dr. V. Gayathri, presented a modified J48 classifier to increase the veracity about the data mining process. They approached two algorithms; the first algorithm is well known as J48 and the other one is J48 with Ant Optimization, as it comes from Ant Colony Optimization (ACO). It is a framework driven specialist that recreates the common conduct of ants. It includes the components of participation and adjustment. This is a new metaheuristic which appears to be both powerful and flexible as it has been effectively connected to a scope of various combinatorial improvement issues. In conclusion, the researcher used weka to test negative (TN) and tested positive (TP) values of the dengue dataset concerning their attributes. The precision of the proposed learning algorithm is 87.52% than others that showed 80% [4].

Tina R. Patil, S.S.Sherekar, describes the Bank dataset for evaluation to maximize true positive and reduce false-positive rate moderately achieving higher classification accuracy using weka. She performed classification using two classifiers, naïve Bayes algorithm, and J48 decision tree. According to "Yes" and "No" values of the attributes the accuracy of J48 gave the more accurate result of 31% for Yes and 87% for No, whereas Naïve gave only 9% for Yes and 89% for No. The cost analysis was the same for both the classifiers [5].

Shameem Fathima, Nisar Hundewale, the researcher took real-time data from specialty hospitals and diagnostic laboratories. They used SVM and Naïve Bayes Classifier for data mining techniques and also used proficient methodology - random classifier with associated Gini features. They also had a classifier with a random forest. SVM with random forest achieved the best accuracy of 0.9078 while naïve Bayes is more sensitive than SVM. So, the conclusion is that SVM, when combined with random forests, has an accurate diagnostic ability [6].

Thypparampil Karunakaran Sajana, Maroju Navya, YVSSV. Gayathri, Nimgire Reshma, applied non-invasive machine learning techniques to comfort the doctors who ordered the hazards in dengue patients. They conducted a corresponding study among Simple Classification Regression Tree (CART), C4.5 algorithms, and Multilayer perception. Definitely, the CART algorithm showed 100% accuracy for the classification of affected or unaffected patients. A dengue dataset was collected for the confusion matrix. The performance metrics of algorithms based on the classification ratio and confusion matrix evaluated Simple CART which showed the best results out of all with a 100% accuracy rate and 1 % of precision and recall of 20 patients samples of dengue fever [7].

Nandini. V, Sriranjitha. R, Yazhini. T. P, aimed at operating named body recognition to extract disorder mentions. That can be recycled to conclude the presence of dengue. They then performed a frequency analysis that correlates the circumstances of dengue along with the symptoms over months. This produces accuracy and serves as a valuable tool for medical experts. The conclusion of this analysis is the detailed design and related algorithms (Lib SVM, Logistic, MLP, NB, SMO, and Simple Logistic) to identify disorder mentions from text and correlated its frequency with the time. The annotated clearance analysis is tagged and aspect extraction finding algorithms are used to achieve the countenance relevant to dengue. This is followed by the generation of a component vector. The Binary representation is then worn towards the train and form various classification representation model and SMO (Sequential Minimal Optimizer) is found to be 91% accurate and generate outstanding results. This model produces further benefits in the forecast of the disease. Additionally, the correlation of instruction samples with time was related to the interaction obtained from predicted consequence and the dengue occurrence abide found to fixate in August, September, and October [8].

Nelofar Rehman described that big data consists of large volumes of complex datasets. And these big data are now used by IT professional engineers and researchers for working. So, to determine the problem of big data he chose data mining tools like Weka, Rapid miner, Mahout, Orange, and Data Melt. He showed the techniques which are used for the best performance of analysis are Classification, Clustering, and Regression. All of these techniques with their related algorithms c4.5, SVM, K-Mean, BIRCH, and CURE are used to determine the Business problem [9].

Kashish Ara Shakil, Samiya Khan, Shadma Anis, Mansaf Alam, used WEKA with 10 cross-validations to evaluate the dengue dataset and compared the different data mining approaches in weka over explorer, knowledge flow and Experiment interface. To determine prediction, they took 108 instances and 18 attributes, their accuracy, and the classification of different algorithms. For the extraction of appropriate information against data, they applied Naïve Bayes, J48, RANDOM tree, SMO and REP tree. But they concluded that J48 and Naïve Bayes are the best performing algorithms for classification. They achieved 100% accuracy with 99 correct instances, ROC=1, with least mean absolute error and it took minimum time for explorer and Knowledge flow to produce results [10].

Iqra Jahangir, Abdul-Basit, Abdul Hannan, Sameen Javed, prepared association rule mining for prediction of dengue. They applied the patient's data in the weka by using the Apriori algorithm. They used explorer from weka for prediction and the rules were generated which evaluated the accuracy using these techniques. The result generated from the rules of apriori was compared along with the earliest result on the basis of accuracy. It concluded that the identification of correct records had an accuracy of 75% [11].

R. Sanjudevi, D.Savitha, applied four phases for improving the prediction accuracy of dengue disease. In the first phase, they collected data of dengue disease from the UCI repository. In the second aspect, the selection is done by the forwarding and backward stepwise regression method. In the third phase, they used SVM and DT (Decision Tree) algorithms and applied them to the dataset. In the last phase, they determined the accuracy by calculating Sensitivity, AUC and Specificity. Finally, it was concluded that SVM is the top performance classifier technique, and achieved an accuracy of 99% and takes fewer times to run in less error rate [12].

M. Bhavani, S.Vinod Kumar, described calculating the work of assorted classification Techniques. They used five techniques REP Tree, SMO, ZeroR, J48, and Random Tree. The Dengue Dataset was analyzed, and a comparison was done between these techniques. Weka tool measured TP rate, correctly classified, FP rate, incorrectly classified and precision through the classifier. This concludes that SMO and J48 was well-performed comparison. They achieved 84% and 76% accuracy rate respectively [13].

Kamran Shaukat, Nayyer Masood, Sundas Mehreen, Ulya Azmeen, collected the dataset from DHQ Jhelum. They evaluated techniques independently with the aid of tables and charts calculated on the dengue datasets. They did experiments in weka. Weka concludes that Naïve Bayes is greatest among all others. Naïve Bayes gives the best accuracy of 92% while RT and REP tree didn't give us probability [14].

III. METHODOLOGY

Weka (Waikato Environment for Knowledge Analysis) software is a free available machine learning tool used as the data mining tool. It is written in a suitable language for portability that is Java Programming Language and it is developed by the Waikato University of Hamilton, New Zealand. Weka offers a set of algorithms for mining that we can apply directly to the arff text file. ARFF (Attribute-Relation File Format) files were also advanced by the Machine Learning Program found in the Department of CS of The Waikato University with the Weka software. It is the file of ASCII text that particularly specifies the details of instances allocation of attributes. In this research, we have classified the accuracy of the dengue dataset. Bioinformatics is approached by different algorithms to classify the disease dengue. Weka has interfaces such as Explorer, Experimenter, Knowledge flow, Workbench and Simple CLI which we are going to use in this report.

1) Explorer: It has different panels (a) Preprocess, (b) Classify, (c) Cluster, (d) Associate, (e) Select attributes, and (f) Visualize. But in all this, our main element is the Classification Panel.
2) Experimenter: In this block, it facilitates precise observation of different algorithms on given datasets. Every algorithm runs ten times and then the correctness is shown.
3) Knowledge Flow: An alternative of the explorer block is the Knowledge flow. There is one difference is that the users select the weka component from the toolbar and attach them to prepare a layout.
4) Workbench: Makes it very easy to complete interactive experiments, so it is unexpected that most work has been finished with small to medium-sized datasets. However, broad datasets have been well processed.
5) Simple Command-line interface: It is a procedure performed through a cmd-line interface by offering instructions to the OS. This interface is least popular compared to others.

1. CLASSIFICATION

Classification is the function of objects, for one of a few predefined categories. It is a widespread issue that covers many diverse applications. This includes locating, classifying, and predicting, etc. A classifier can be used to predict class labels of unknown records. It typically uses a training set where all objects are already associated with known classes and a training set containing records whose class is known. The evaluation of the classifier is based on the calculation of records accurately or inaccurately predicted by the instrument.

This technique is a systematic approach to construct models from input datasets. Classification techniques are more favorable for analyzing datasets with binary or nominal ranges. These techniques include Naïve Bayes, SMO, ZeroR, J48, REP Tree, Random Tree, etc. These techniques are employing a learning algorithm to select a model that best fits the bond between the attribute set and the class label with an input dataset. It is an algorithm with good capability i.e., that correctly predicts the class of unknown records.

In this paper, the technique uses the explorer interface and depends on related algorithms. Classification is the main part of our prediction of dengue datasets. We take cross-validation for better accuracy. We select the training sample and testing sample from the model and then classify the accuracy for each classifier. A classifier has many classifications to analyze the perfect result like Correctly Classified Instances, Incorrectly Classified Instances, and Kappa statistic, mean absolute error, root mean squared error, relative absolute error, and Root squared error and Total Number of Instances.

2. DATASETS USED

Dataset is a very useful part of the research for predicting. It is a collection of massive amounts of data. The Dengue dataset is tweets collected from twitter through the Vicinitas tool. It stores data in an excel file or .xls format of the dataset with many attributes related to the twitter users. A classifier can be treated as a black box that naturally assigns a class when presented along with the attribute set of an undiscovered record. It collects data of a year, after which, we need only some meaningful attributes from the file. So, it has filters related to our need and converted into arff files of the dataset which is taken by the weka tool. We have a dataset of tweets with hashtags dengue and keyword dengue on Twitter.

Table 1 Attributes of Dataset

Sentence	Class
I had another attacked of dengue but had got nearly over it	No
Dengue is spreading everywhere to be careful.	Yes
Make sure you're aware of mosquito bites in this monsoon.	Yes
I had dengue but I’m fine now.	No

We have taken the dataset that has #dengue and the keyword “dengue” or normal sentence including the word “dengue”. As shown in Table 1 we have divided it according to grammatical tenses, i.e. past tense and present tense. Generally, users tweet in the present tense, so we know that the incident is happening right now. If the user wants to write about his past, then he/she tweets in the past tense, so we have divided the keyword dengue into the past tense and present tense by using classes ‘Yes’ and ‘No’.

The datasets have a Total of 102 tweets in which we have applied 65 True or Yes conditions and 35 False or No conditions. They support the arff file, so files have two attributes in the program. The first attribute is the text with “string” data type and the second is divided into ‘Yes’ or ‘No’ class respectively.

3. DATA MINING TECHNIQUES

This paper is only focused on the Explorer Interface of Weka. In this procedure, the dataset is loaded in (arff) format files. We classified dataset into a string to vector form which shows one by one attribute in the class nominal.

There are Naïve Bayes, SMO, ZeroR, J48, and Random Forest, Random Tree, and REP Tree data mining techniques for predicting dengue disease. The interface processes the data and filter string into vector form after which we apply the following classifiers.

3.1. NAÏVE BAYES CLASSIFIER:

Generally, NB uses the Bayes formula i.e. it performs arithmetic progression. Individually classifying the data attributes as a probability classifier i.e. analyzing the classifier output with statistic-based o/p by using cross-validation [10]. A simple Naïve classifier ensures comparable performance with a neural system classifier.

After running this classifier for 0.03 sec, we got an o/p of 90 correctly classified instances with 88.2523% of accuracy and least mean absolute error of 0.1454, Matthews Correlation Coefficient (MCC) of 0.743, Receiver Operating Characteristics (ROC) of 0.911 and Precision-recall Curve Area (PRC) of 0.917.

3.2. SMO CLASSIFIER:

Sequential minimal optimization (SMO) is an algorithm for solving the quadratic programming (QP) problem that arises during the training of SVM. It is used for SVM and is implemented by the popular LIBSVM tool. Platt's SMO algorithm is well organized with good computational efficiency. This technique is applied to a dataset for splitting our data.

After running this classifier, we got the output of the classifier by altered measurements to create an indication for every occurrence of the Dengue dataset. The output produced by SMO achieved an accuracy of 86.2745% with the PRC Area of 0.808 in 0.03 seconds.

.

3.3. ZEROR CLASSIFIER:

ZeroR is the sorting method that depends on the target. The ZeroR simply gives a prediction of the majority class (categories) and ignores all predictors. However, ZeroR does not have any estimation power, which is suitable for deciding criterion performance as a reference point for other classification methods.

We got outcomes after evaluation in the weka Tool as a result. Based on 10 cross-validations to each instance, 63.7255% accuracy was achieved in 0 seconds with the least ROC Area of 0.452.

3.4. J48 CLASSIFIER:

C4.5 (J48) is used to make a decision tree. The J48 is used for classification and it is also referred to as a numerical statistical classifier. C4.5 is an addition to the Iterative Dichotomiser 3 algorithm. ID3 Technique constructs a decision tree from a set of data, with the Conception of information entropy. To decide the value based on attributes of the dataset to foreshow and classify the accuracy we perform the J48 classification.

To predict using machine learning we ran this classifier and we got 88 accurate instances with 86.2745% of accuracy in 0.13 seconds. While 10 cross-validations test it gives 0.1759 mean value error rate.

3.5. RANDOM FOREST CLASSIFIER:

Random Decision Forest learning manner is used for task prediction and classification by constructing a decision tree. This method is created with the subspace formula method for approaching classification. Recoil is also done by this classifier with the dataset.

To calibrate correct classification we are using a cross-validation option which is optimum for massive datasets. After performing this classification we got status within 0.5 seconds, with the accuracy of 75.7902% and the error rate of 0.3538 to acquire their best result with 0.084 of ROC Area.

3.6. RANDOM TREE CLASSIFIER:

Random Tree (RT) is a learning algorithm that is an organized classifier. It is an algorithm for constructing a tree that chooses attributes at each node. Each node is echeloned using the best variables to allow the class probabilities. For producing a random set of information data to construct a decision tree.

We analyzed the output with the statistic-based method by using 10 cross-validations for the prediction of the dengue dataset. We got an accuracy of 68.6275% with a mean absolute error of 0.3137 and occupied 0.655 of the ROC Area.

3.7. REP TREE CLASSIFIER:

REP Tree (Decision/Regression tree) has been used to build a decision/regression tree using entropy as an impurity measure. It uses reduced-error pruning by sorted values for numeric property and fission the instances into pieces to classify the accuracy. It creates several trees in different iterations, after that it picks one from all trees.

We tested our dataset with REP Tree and we got the outcomes with 88.2353% of accuracy and 90 correctly classified instances in 0.06 seconds. The Mean absolute error of this algorithm is 0.208 and the precision-recall curve area shown is 0.795 under the class.

IV. RESULTS AND DISCUSSION

The data mining techniques that we performed used seven different algorithms i.e. Naïve Bayes, SMO, ZeroR, J48, Random Forest, Random Tree, and REP Tree. In Figure 1, we can see that the bar graph of all the algorithms according to their accuracy rate classified as correct instances, the area covered by algorithm, the correlation between them and error rate. We found out that Naïve Bayes is the best algorithm as compared to others.

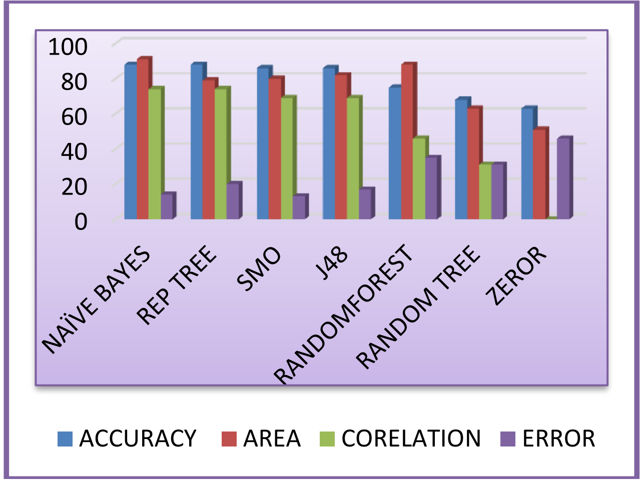

Figure 1: Bar Graph of Algorithms applied in dengue dataset

Table 2 shows Naïve Bayes and REP Tree classifications, with true positive rates of 88.2353%, which is the maximum accurate rate for predicting dengue instances using grammatical tense form. But both having different mean absolute error i.e., NB having 0.14 Mean Absolute Error while REP Tree having an error rate of 0.20 which means naïve is a good algorithm with less error rate. Naïve Bayes has the maximum correlation value with a true positive rate and MCC (Matthews correlation coefficient) of 0.74, which is the highest measurement compared to other algorithms and the precision-curve area is maximum in Naïve Bayes (PRC=0.917) which shows the prediction of disease is maximum with ROC=0.9 rate.

Secondly, SMO and J48 algorithms have an accuracy rate of 86.2745% which is close to the maximum rate. Then the others like Random Forest, ZeroR and Random Tree did not give a sensitive result. To predict the survivability of dengue disease with the best performance, the Naïve Bayes algorithm is highly preferred.

Table 2 List of Algorithms Results

Algorithms	NAÏVE BAYES	SMO	ZEROR	J48	RANDOM FOREST	RANDOM TREE	REP TREE
Attributes							
Correctly Classified Instances	88.2353%	86.2745%	63.7255%	86.2745%	75.4902%	68.6275%	88.2353%
Incorrectly Classified Instances	11.7647%	13.7255%	36.2745%	13.7255%	24.5098%	31.3725%	11.7647%
Mean Absolute Error	0.1454	0.1373	0.4635	0.1759	0.3538	0.3137	0.208
Precision	0.882	0.862	0.637	0.863	0.784	0.683	0.888
F-measure	0.882	0.861	0.778	0.860	0.723	0.684	0.879
MCC	0.743	0.699	0.0	0.698	0.460	0.314	0.745
ROC Area	0.911	0.840	0.452	0.834	0.884	0.655	0.786
PRC Area	0.917	0.808	0.515	0.828	0.886	0.632	0.795

V. CONCLUSION AND FUTURE SCOPE

In this paper, our main objective is to use tweets from twitter for analyzing an arboviral disease “dengue” prediction. We focused on tweets to find out how fast the dengue spread than the government surveillance is and how much the results should be correct for dengue to be trending on twitter. We came to the conclusion that Naïve Bayes is a good performer for classification analysis, whereas Naïve Bayes and REP Tree have the highest accuracy rate where Naïve has the least mean absolute error.

So, we conclude that the best outcomes are achieved by Naïve Bayes having the highest accuracy rate of 88.23% which shows that the prediction of dengue in twitter is easily done by this algorithm.

Therefore, in this research, we found out that twitter with a huge number of tweets can be easily used for the prediction of dengue disease. As well as we can say that hashtag dengue (#dengue) or dengue tweets are helpful for the prediction of dengue disease. Our future work includes a prediction for other diseases using tweets that are classified as text mining. Through this, in India, people can be more aware of the disease and its dangers and can start taking preventive measures and precautions against dengue at the earliest.

ACKNOWLEDGMENT

I would like to thank my advisor Dr.Klinsega Jeberson and Dr.Wilson Jeberson for their guidance and support in assisting me with the entire research. They have been with me through high and low moments I experienced during research. I am very grateful to the faculty of Computer Science and may God bless them. Special acknowledgment to my brother Mr. Sanjay Kumar and friends who supported me, when I was started this degree. Finally, I am very thankful to my whole family. They are like my backbone especially to my mother Mrs. Shobha Kumari and father Mr. Santosh Kumar for their support and encouragement throughout my studies.

REFERENCES

[1] Andrea villanes, Emily Giffiths, Michael Rappa, Christopher G. Healey, "*Dengue fever surveillance in India using text mining in public media*", The American Journal of Tropical Medicine & Hygiene, Vol.**98,** Issue.**1,** pp.**181-191**, **2018**.

[2] Wajeeha Farooqi, Sadaf Ali, "*A critical study of selected classification algorithms for dengue fever and dengue hemorrhagic fever*", In the Proceedings of the 2013 IEEE 11th International conference on Frontiers of Information Technology (FIT 2013), **USA**, pp.**140-145**, **2013.**

[3] M. Vidhyalakshmi, P. Radha*, "Social HashTag Techniques Using Data Mining-A Survey"* International Journal of Scientific Research in Computer Science and Engineering, Vol.**6**, Issue.**3**, pp.**86-92**, **2018**

[4] N.Saravanan, Dr. V. Gayathri, "*Classification of dengue dataset using J48 algorithm and ant colony based a J48 algorithm*", In the Proceeding of the 2017 International Conference on Inventive Computing and Informatics (ICICI 2017), **India**, pp.**1062-1067**, **2017**.

[5] Tina R.Patil, S.S.Sherekar, "*Performance analysis of naive Bayes and j48 classification for data classification*", International journal of computer science and applications, Vol.**6,** No.**2**, pp.**256-267**, **2013.**

[6] Shameem Fathima, Nisar Hundewale, "*Comparison of classification techniques-SVM and naive Bayes to predict the arboviral disease-dengue*", In the Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), Atlanta, **GA**, pp.**538-539**, **2011**.

[7] Thypparampil Karunakaran Sajana, MarojuNavya, YVSSV.Gayathri, Nimgire Reshma, "*Classification of dengue using machine learning techniques*", International Journal of Engineering and Technology, Vol.**7**, Issue.**2.32,** pp.**212-218**, **2018**.

[8] Nandini. V, Sriranjitha. R, Yazhini. T. P, "*Dengue detection and prediction system using data mining with frequency analysis*", 6th International Conference on Advances in Computing and Information Technology, pp.**53-67, 2016**.

[9] Nelofar Rehman, "*Data Mining Techniques Method Algorithms and Tools*", International Journal of Computer Science and Mobile Computing, Vol.**6**, Issue.**7**, pp.**227-231**, **2017**.

[10] Kashish Ara Shakil, Samiya Khan, Shadma Anis, Mansafalam, "*Dengue disease prediction using weka data mining tool*", In the Proceeding of IIRAJ International Conference (ICCI-SEM-2K17), **India**, pp.**48-59**, **2015**.

[11] Iqra Jahangir, Abdul-Basit, Abdul Hannan, Sameen Javed, "*Prediction of dengue disease through data mining by using modified apriori algorithm*", In the Proceeding of ACM 4th International Conference of Computing for Engineering and Sciences (ICCES 2018), **Malaysia**, pp.**1-4**, **2018**.

[12] R. Sanjudevi, D.Savitha, "*Dengue Fever Prediction Using Classification Techniques*", International Research Journal of Engineering and Technology (IRJET 2018), Vol.**6**, Issue.**2**, pp.**558-563**, **2018**.

[13] M.Bhavani, S.Vinod Kumar, "*A data mining approach for precise diagnosis of dengue fever*", International Journal of Latest Trends in Engineering and Technology, Vol.**7**, Issue.**4**, pp. **352-359**, **2016**.

[14] Kamran Shaukat, Nayyer Masood, SundasMehreen, UlyaAzmeen, "*Dengue fever prediction: a data mining problem*", Journal of Data Mining in Genomics and Proteomics, Vol.**6**, Issue.**3**, pp.**1-5**, **2015**.

[15] Amit Palve, Rohini D.Sonawane, Amol D. Potgantwar*, "Sentiment Analysis of Twitter Streaming Data for Recommendation using Apache Spark"* International Journal of Scientific Research in Network Security and Communication, Vol.**5**, Issue**.3**, pp.**99-103**, **2017.**

Author's Profile

Miss. Sarita Kumari pursuing a Master's degree in Computer Science and Engineering from Sam Higginbottom University of Agriculture, Technology & Sciences, Allahabad, India. She had received her Bachelor's degree in Computer Science and Engineering from Sam Higginbottom University of Agriculture, Technology & Sciences, Allahabad, Uttar Pradesh, India.

Dr. Klinsega Jeberson has received her Ph.D. degree in Computer Science and Communication, from Sam Higginbottom University of Agriculture, Technology & Sciences, Allahabad, Uttar Pradesh, India Currently, she is working as a (Sr. Grade) Assistant Professor in the Department of Computer Science & I.T., Sam Higginbottom University of Agriculture, Technology & Sciences, Allahabad, Uttar Pradesh, India. She has published more than 10 research papers. Her area of interest includes Data Mining and Web Technologies.

Dr. Wilson Jeberson was awarded a Ph.D. degree in Computer Science and Communication, from Sam Higginbottom University of Agriculture, Technology & Sciences, Allahabad, Uttar Pradesh, India Currently; he is working as Professor and Head, Department of Computer Science & Information Technology in Sam Higginbottom University of Agriculture, Technology & Sciences, Allahabad, Uttar Pradesh, India from 2003. He has published more than 50 papers in reputed international journals and more than 15 Papers in National & International Proceedings. A research paper was selected as best one among the top 20 papers presented in the Second GMSARN International Conference at Pattaya, Thailand and was published in the International Journal of AIT. He is one of the editorial board members of various journals including Scientific & Academic Publishing Co., Journal Name: Software. Engineering, USA.